AudioGen: Textually Guided Audio Generation

Year: 2022 | Citations: 382 | Authors: F. Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre D'efossez

---

Abstract

We tackle the problem of generating audio samples conditioned on descriptive text captions. In this work, we propose AaudioGen, an auto-regressive generative model that generates audio samples conditioned on text inputs. AudioGen operates on a learnt discrete audio representation. The task of text-to-audio generation poses multiple challenges. Due to the way audio travels through a medium, differentiating ``objects'' can be a difficult task (e.g., separating multiple people simultaneously speaking). This is further complicated by real-world recording conditions (e.g., background noise, reverberation, etc.). Scarce text annotations impose another constraint, limiting the ability to scale models. Finally, modeling high-fidelity audio requires encoding audio at high sampling rate, leading to extremely long sequences. To alleviate the aforementioned challenges we propose an augmentation technique that mixes different audio samples, driving the model to internally learn to separate multiple sources. We curated 10 datasets containing different types of audio and text annotations to handle the scarcity of text-audio data points. For faster inference, we explore the use of multi-stream modeling, allowing the use of shorter sequences while maintaining a similar bitrate and perceptual quality. We apply classifier-free guidance to improve adherence to text. Comparing to the evaluated baselines, AudioGen outperforms over both objective and subjective metrics. Finally, we explore the ability of the proposed method to generate audio continuation conditionally and unconditionally. Samples: https://felixkreuk.github.io/audiogen